

#12

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-287969

(43)Date of publication of application : 31.10.1995

(51)Int.Cl.

G11B 27/28

G10L 3/00

G10L 3/00

G10L 3/00

H04R 3/00

(21)Application number : 07-082900

(71)Applicant : XEROX CORP

(22)Date of filing : 07.04.1995

(72)Inventor : VIJAY BALASUBRAMANIAN

CHEN FRANCINE R

PHILIP A CHOU

DONALD G KIMBER

ALEX D POON

WEBER KARON A

LYNN D WILCOX

(30)Priority

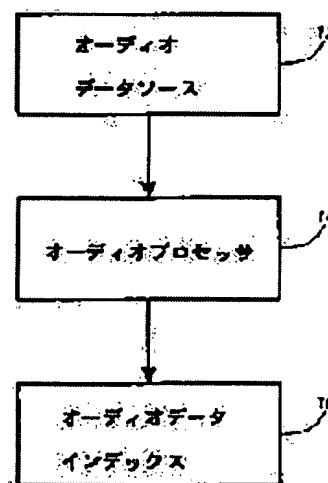
Priority number : 94 226580 Priority date : 12.04.1994 Priority country : US

(54) SYSTEM OF PROCESSOR CONTROL

(57)Abstract:

PURPOSE: To prepare an index in an audio data stream.

CONSTITUTION: An audio stream is given from an audio data source 12, and the data are imparted by a speaker conducting a conversation, a recording video with an audio track or other audio sources. Audio data are transmitted to an audio processor 14, the audio processor is arbitrarily known as a general purpose computer, and the audio processor outputs an audio data index 16.



LEGAL STATUS

[Date of request for examination]

08.04.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平7-287969

(43)公開日 平成7年(1995)10月31日

(51)Int.Cl. ⁹	識別記号	庁内整理番号	F I	技術表示箇所
G 1 1 B 27/28		A 8224-5D		
G 1 0 L 3/00	5 3 1 L			
	5 3 5			
	5 5 1 G			
H 0 4 R 3/00	3 1 0			

審査請求 未請求 請求項の数3 O L (全 13 頁)

(21)出願番号 特願平7-82900

(22)出願日 平成7年(1995)4月7日

(31)優先権主張番号 2 2 6 5 8 0

(32)優先日 1994年4月12日

(33)優先権主張国 米国 (U S)

(71)出願人 590000798

ゼロックス コーポレイション
XEROX CORPORATION
アメリカ合衆国 ニューヨーク州 14644
ロチェスター ゼロックス スクエア
(番地なし)

(72)発明者 ヴィジャイ パラスプラマニアン
アメリカ合衆国 ニュージャージー州
08544プリンストン プリンストン ユニ
バーシティ オールド グラジュエイト
カレッジ ルーム 184

(74)代理人 弁理士 中島 淳 (外1名)

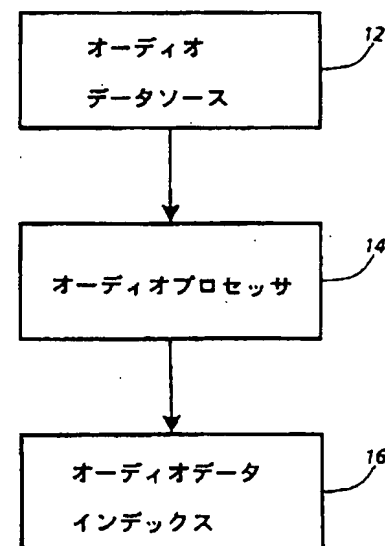
最終頁に続く

(54)【発明の名称】 プロセッサ制御のシステム

(57)【要約】

【目的】 オーディオデータストリーム内にインデックスを作成する。

【構成】 オーディオストリームはオーディオデータソース12から与えられ、該データは、会話を行うスピーカー、オーディオトラックを伴う記録ビデオ、または他のオーディオソースによって与えられることが可能である。オーディオデータはオーディオプロセッサ14へ送られ、オーディオプロセッサは汎用コンピュータのような任意の公知デバイスであることが可能であり、本発明に従って構成されることが可能である。オーディオプロセッサはオーディオデータインデックス16を出力する。



【特許請求の範囲】

【請求項1】 リアルタイムに記録されるオーディオデータに対してスピーカに從う電子インデックスを相関付けるプロセッサ制御のシステムであって、前記オーディオデータは複数の個々のスピーカからの音声を含み、前記システムは、

複数の個々のスピーカの各々に対するスペクトル特徴トレーニングデータを与えるトレーニングデータソースと、

トレーニングデータを受信して複数の個々のスピーカの各々に対するスピーカモデルを生成するシステムプロセッサであって、各スピーカモデルは関連するスピーカ識別子を有し、前記システムプロセッサはさらに前記スピーカモデルを結合してスピーカネットワークとする、システムプロセッサと、

複数の個々のスピーカからの音声を含むリアルタイムオーディオデータを与えるオーディオ入力システムと、前記オーディオデータを受信して前記オーディオデータをスペクトル特徴データに変換するオーディオプロセッサと、

前記オーディオデータを受信して受信時間に従って記憶媒体上へ前記オーディオデータを記録する記録デバイスと、

データを記憶するメモリであって、メモリに記憶されるデータはプロセッサが実行する命令を示す命令データを含む、メモリと、

前記システムプロセッサはさらに、メモリに記憶されたデータにアクセスし、

前記システムプロセッサは、命令実行の際に、前記スペクトル特徴データを前記オーディオプロセッサから受信し、前記スピーカネットワークを使用して、異なる個々のスピーカモデルに対応する前記オーディオデータのセグメントを決定し、

前記システムプロセッサはさらに、各セグメントの開始においてタイムスタンプを決定し、前記タイムスタンプは前記記憶媒体上の当該セグメントに対する受信時間に対応し、前記システムプロセッサは前記タイムスタンプを前記メモリに記憶し、

前記システムプロセッサはさらに、当該セグメントに対する前記記憶媒体のロケーションアドレスと共に、各セグメントに対する前記個々のスピーカモデルの前記スピーカ識別子を前記メモリに記憶する、プロセッサ制御のシステム。

【請求項2】 前記システムプロセッサは複数の個々のスピーカの各々に対して個々のHMMスピーカモデルを生成する、請求項1に記載のシステム。

【請求項3】 前記システムプロセッサはさらに前記個々のHMMスピーカモデルを並列に結合してスピーカネットワークHMMを形成する、請求項2に記載のシ

ステム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は記録またはリアルタイムオーディオデータストリームに対する会話音声のスピーカ（話し手）によるセグメンテーションに関する。

【0002】 より詳細には、本発明は既知のスピーカを含むオーディオデータのリアルタイム記録期間中に会話音声を自動的にセグメンテーションするシステムに関する。

【0003】

【従来の技術】 オーディオおよびビデオ記録は、コンシューマグレード（消費者レベル）の記録装置の発展によって今や一般のものとなっている。後の再生のための過去の記録としてビジネスミーティング、講義、もしくはバースデーパーティーが記録されることは今や稀なことではない。不幸にして、オーディオおよびビデオ媒体の両者は、所望の記録部分にアクセスする際のアシストとなる外部またはオーディオ情報をほとんど与えない。書籍においては、巻頭の目次および巻末の索引によってインデックス化が与えられ、このインデックス化によって読者は複数の著者の確認および複数の著者の参照を容易に行うことが可能である。同様のインデックス化方法がオーディオストリームにおいて有用であり、ユーザーは特定のスピーカの会話部分を確認することが可能となる。ほとんどのビデオ記録に関連する限られたデータ量は、見る者が確実におよび容易に所望の関心部分にアクセスするための十分な情報を与えない。このため見る者は記録内容を順に調べて所望の情報を検索しなければならない。

【0004】 例えばスピーカ（話し手）やトピック（主題）を示すノートのような、記録中に取られたノートが検索の補助となることが可能である。このようなノートは構造的アウトラインを与えるが、ビデオ媒体とノート媒体との間には直接的な相関がないため、ノートの内容を共にしたビデオ上の時刻の補完を強いられる。このことは、非相関媒体におけるイベントノートは通常イベントの継続時間を含まないという事実によって複雑化する。加えて、そのようなノート化またはインデックス化は非常に煩わしい。コンピュータシステムがイベント期間中のノート取得に使用されることが可能であり、該システムは同時に記録されるかまたは事前に記録される。キーボードを使用するテキストベースシステムがこの場合に使用されることが可能であるが、ほとんどの人はタイプするよりもかなり速く話すため、内容を記述するコンピュータ生成テキストラベルをリアルタイムで作成することは相当な努力を必要とする。

【0005】

【発明が解決しようとする課題】 オーディオストリームにおいて異なるスピーカを示すスピーカチェンジマ

3

一カーは、異なるシーケンシャルデータへのランダムアクセスを可能とする。リアルタイム設定においては、そのようなオーディオセグメンテーションは、記録が行われている時にその記録の中へ有用なインデックスを作成する際の補助となり得る。各セグメントは1個人による発声を表す。同一のスピーカーによる発声は結合され、また同様に参照されてインデックスが形成される。会話におけるポーズまたは沈黙期間もまたオーディオインデックス形成において重要である。

【0006】オーディオストリーム内にインデックスを作成することは、リアルタイムであっても処理後であっても、ユーザーが特定のオーディオデータセグメントを認識することを可能にする。例えばこのことは、ユーザーが記録を拾い読みして特定のスピーカーに対応するオーディオセグメントを選択したり、次のスピーカーへ記録を早送りすることを可能にする。加えて、スピーカーの順序を知ることが、会話または会話の内容に関する内容情報を与えることも可能である。

【0007】

【課題を解決するための手段】本発明は、リアルタイムに記録されるオーディオデータに対してスピーカーに従う電子インデックスを関連付けるプロセッサ制御のシステムを与える。該システムは、複数の個々のスピーカーの各々に対するトレーニングデータソースを含む。オーディオ入力システムは、個々のスピーカーに対する音声を含むリアルタイムオーディオデータを与える。オーディオデータはオーディオプロセッサによってスペクトル特徴データに変換されると同時に記録デバイスによって記憶媒体上に記録される。システムプロセッサはトレーニングデータを受信して個々のスピーカーモデルを作成し、該モデルは並列に接続されてスピーカーネットワークが形成される。システムプロセッサは次にオーディオデータのスペクトル特徴データを受信し、スピーカーネットワークを使用して各スピーカーに対応するオーディオデータ内のセグメントを決定する。

【0008】

【実施例】図1は一般化されたオーディオ処理システム10のブロック図を示し、該システムにおいて本発明が実施されることが可能である。一般に、オーディオストリームはオーディオデータソース12から与えられ、該データは、会話を行うスピーカー、オーディオトラックを伴う記録ビデオ、または他のオーディオソースによって与えられることが可能である。オーディオデータはオーディオプロセッサ14へ送られ、オーディオプロセッサは汎用コンピュータのような任意の公知デバイスであることが可能であり、本発明に従って構成されることが可能である。オーディオプロセッサはオーディオデータインデックス16を出力する。

【0009】図2はオーディオインデックスシステムの一般化されたフロー図を示す。図2に示されるステップ

4

は以下により詳細に説明されるが、図2は本発明により記述される方法の概観を与えるものである。

【0010】オーディオ波形20はボックス22のステップにおける入力である。ボックス22におけるオーディオストリームは、処理されるべきオーディオの部分を含むことが可能であるが、オーディオストリーム内の全てのスピーカーからの音声を含まなければならない。説明を目的として、オーディオストリーム全体がボックス22のステップにおける入力である。ボックス24のステップは音声信号データをスペクトル特徴ベクトルへ変換する。例えば、12次のケプストラムが20msごとに算出されることが可能である。

【0011】ボックス26のステップにおいて、HMMスピーカーモデルは初期化データに基づき各スピーカーに対してトレーニングされる。複数の個々のスピーカーモデルは、モデルを並列に接続することによってボックス28のステップにおいて結合され、会話のHMMスピーカーモデルが形成される。

【0012】ボックス30のステップは、セグメンテーションが実行されるオーディオストリームを入力する。オーディオストリームはボックス22のステップで使用されるトレーニングオーディオデータを含んでも含まなくてもよい。スピーカーモデルの事前トレーニングに対してスピーカーが使用可能である場合、入力されるオーディオストリームもまたリアルタイムに発生およびセグメンテーションされることが可能である。ボックス32のステップにおいて、入力されるオーディオから再び特徴が抽出され、この特徴抽出はボックス24のステップにおけるものと同様である。

【0013】ボックス34のステップはボックス28のHMMスピーカーネットワークを使用し、入力されるオーディオストリームのセグメンテーションを行う。セグメンテーションはビタビ(Viterbi)デコーディングを使用して行われ、スピーカーネットワークを介する最も確からしい状態シーケンスが見出され、状態パスがスピーカーを変更する場合にはマーキングが施される。

【0014】セグメンテーションとインデックス化の確度は、ボックス26のステップに戻ってスピーカーモデルを再トレーニングすることによる後処理の適用で改善されることが可能であり、この場合ボックス34のステップからのセグメンテーション情報が使用される。一般に、セグメンテーション情報に基づき、スピーカー当たりより多くの音声情報が使用可能となり、このことはより詳細なスピーカーモデルが決定されることを可能とする。再トレーニングと再セグメンテーションの繰り返しは、ボックス34のステップでのセグメンテーションで大きな変化が生じなくなるまで続けられることが可能である。ボックス32のステップにおける特徴抽出の結果はまたセーブされることが可能であり、各再トレーニングの繰り返しと共に再使用されてボックス34のステッ

5

ブでオーディオデータが再セグメンテーションされることが可能である。

【0015】隠れマルコフモデル(HMM)によるモデル化は音声認識で一般的に使用される統計的方法であり、ワード全体、もしくは単音のようなサブワードがモデル化される。未知の発声の認識は、その発声が最も確からしく与えられるモデルもしくはモデルのシーケンスを見出すことに基づいている。HMMはスピーカ-の識別においても使用されることが可能である。モデルはスピーカ-の発音に対して作成され、その場合発音は特定のワードについてののものであっても自然な音声についてののものであってもよい。スピーカ-の識別は、未知の発声が最も確からしく与えられるスピーカ-モデルを見出すことによって行われる。未知の発声が複数のスピーカ-からの音声を含む場合、スピーカ-は最も確からしいスピーカ-モデルのシーケンスを見出すことによって識別される。

【0016】理論的に、HMMは状態のシーケンスから成り、該状態シーケンスは定められた時間間隔で状態間に発生する遷移を伴う。ある状態への遷移が行われるたびに、その状態の出力特性が発生される。音声認識およびスピーカ-識別の両者において、これらの出力はその時間間隔に対する音声のスペクトル推定を表す。例えばケプストラムがその例である。ケプストラムはスペクトルエンベロップ(包絡線)の推定であり、音声認識およびスピーカ-識別で一般に使用される。ケプストラムは、スペクトルの対数のフーリエ逆変換であり、スペクトルエンベロップと周期的音声ソースとを分離するよう作用する。

【0017】状態間の遷移は出力のシーケンスを特定することによって、HMMが使用されて音声を統計的にモデル化することが可能となる。システムの出力のみが観測されるため「隠れ(hidden)」という言葉が用いられる。即ち、基礎となる状態シーケンスは推定され得るのみである。

【0018】より形式的には、HMM L は、 $S_0 \dots S_{N-1}$ の N 個の状態、状態 i から状態 j への遷移確率 a_{ij} , $i=0 \dots N-1, j=0 \dots N-1$ 、ならびに状態 i で出力 x を生じる確率を与える確率分布 $b_i(x)$, $i=0 \dots N-1$ 、から成る。例えば、 $b_i(x)$ は特徴ベクトル x に対する多変数ガウス分布であることが可能である。加えて、遷移可能であるが出力を発生しないヌル状態が存在する。図3は5状態のHMMを示す。状態 S_0 から状態 S_1 、 S_2 または S_3 への遷移確率は画一的であり、即ち、 $a_{0j}=1/3, j=1,2,3$ である。状態 S_i , $i=1,2,3$ については、自己遷移および状態 S_4 への遷移が存在し、それらは等確率である。従って $a_{ii}=1/2$ および $a_{i4}=1/2, i=1,2,3$ である。状態 S_4 については遷移は常に S_0 へ行われ、従って $a_{40}=1$ である。状態 S_1 、 S_2 、および S_3 に関連す

6

る出力分布は、それぞれ $b_1(x)$ 、 $b_2(x)$ 、および $b_3(x)$ である。状態 S_0 および S_4 はヌル状態であり、従って関連する出力を有さない。状態 S_0 と S_4 を結合することによって等価なHMMが形成されることがかのである。しかし、HMMを結合してより大きなHMMネットワークを形成するタスクを簡素化するために、このことは行われない。これについては以下に説明が行われる。HMMに関するより深い検討は、Rabiner による「A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition」(Proc. IEEE, vol. 1.77, No. 2, February, 1989, pp. 257-285)に見出される。

【0019】対象物のシーケンスをモデル化するネットワークHMMは、以下のように個々のHMMを並列に結合することにより作成される。認識される L 個の対象物の各々に対するHMMを $L_i, i=1, \dots, M$ とする。先に述べたように、対象物は単語、単音、またはスピーカ-のいづれであってもよい。ネットワークHMMは、許容される全ての対象物シーケンスに対して対象物HMM間の遷移を付加することにより作成される。図4において、HMM L_1 、 L_2 、および L_3 によって3つの対象物がモデル化されている。これら対象物は、遷移により示されるように任意の順序で発生可能である。状態 S_0 はヌル状態であり、従って出力を発生しない。 S_0 から、対象物HMM L_1 、 L_2 、および L_3 への遷移は等確率となる。全ての対象物HMMからの遷移は状態 S_0 に向かい、次に状態 S_0 への遷移となる。

【0020】 T 個の出力 $X = x_1 \dots x_T$ のシーケンスが与えられる場合、どの対象物HMMシーケンスが最も確からしく出力シーケンス X を発生したかを決定することにより認識が実行される。これにはビタビアルゴリズムが使用され、最も確からしく出力 X を発生したネットワークを介する状態シーケンスが見出される。シーケンス内の各状態は、認識される対象物の内の1つのHMMに対して特定されるため、最も確からしい状態シーケンスは認識対象物のシーケンスを特定する。図5はビタビアルゴリズムの結果を概略的に示す。 x 軸は時間を示し、 y 軸はネットワークHMM内の現行状態を示す。HMM L_1 、 L_2 、および L_3 に対応する状態は y 軸上の領域によって示される。与えられた出力を結果としてもたらし得る状態シーケンスが多数存在可能であるが、ビタビアルゴリズムは最も確からしい状態シーケンスを見出す。図5はビタビパスを示す。時刻 t_0 において最も確からしい対象物は L_1 である。時刻 t_1 において対象物は L_2 であり、 t_2 においては L_3 である。時刻 t_3 において最も確からしい対象物は L_1 となる。

【0021】HMMに対するパラメータは、次に、遷移確率 a_{ij} および出力確率 $b_i(x)$ である。これらパラメータは、HMMによってモデル化された対象物によって既に発生されたことがわかっている出力 X を用いてHMMをトレーニングすることにより学習されることが可能

である。Baum-Welchプロシジャーとして知られているアルゴリズムが一般に使用される。このアルゴリズムは、トレーニングデータ X の尤度を最大にするパラメータ値を繰り返し処理により見出すアルゴリズムである。該アルゴリズムは、パラメータの初期推定から開始する。続いて以下のステップが実行される。(1) トレーニングデータに基づき、状態間遷移確率および状態からの出力確率を算出する。(2) これらの確率を使用し、遷移確率 a_{ij} および出力確率 $b_i(x)$ の推定値を算出する。ステップ(1)および(2)は収束が得られるまで繰り返される。

【0022】前述のように、隠れマルコフモデルが使用されてスピーカー識別を目的として個々のスピーカーがモデル化されることが可能である。図6に示されるように、(特定の発声に対向する)個々の発声スタイルが35状態HMM60を使用してモデル化されることが可能である。状態 S_0 はヌル状態であり、出力を発生する状態 S_1, \dots, S_{32} および S_{31L} への遷移を伴う。これらの遷移確率は p_1, \dots, p_{32} および p_{31L} により与えられる。これら出力発生状態の各々は、確率 q_1 を伴う自己遷移、ならびに確率 $1 - q_1$ を伴う最終ヌル状態 S_{34} への遷移を有している。ヌル状態 S_{34} は確率1で初期ヌル状態 S_0 へ遷移する。各非ヌル状態はガウシアン出力分布を有しており、平均ベクトルおよび対角共分散マトリックスにより特性付けられる。

【0023】図7はサイレンス(無音)サブネットワークを示す。該サブネットワークは直列に接続された3状態から成る。各状態は、通常もしくは結合されたガウシアン出力分布を有し、該分布はラベル S_{1L} で示されている。この出力分布はまた、スピーカーモデル60のサイレンス状態62における出力分布と同一であり、該分布は状態ラベル S_{1L} で示されている。サイレンスサブネットワークは長時間間隔の無音状態をモデル化するが、会話の発声におけるポーズや短時間間隔の無音状態に対しては適切でない。これらポーズや短時間間隔の無音状態は、個々のスピーカーモデルにおけるサイレンス状態62によってモデル化される。スピーカーHMMのサイレンス状態における出力分布は全て結合されてサイレンスサブネットワークにおける出力分布となる。

【0024】スピーカーHMMの各々は、与えられたスピーカーの発声スタイルに対してトレーニングされなければならない。このトレーニングは先に述べたBaum-Welchアルゴリズムを使用して行われ、遷移確率 a_{ij} 、およびガウシアン出力確率 $b_i(x)$ に対する平均および対角共分散が推定される。HMMパラメータの初期推定値は次のように得られる。全ての遷移確率が画一的に設定され、この結果、与えられた状態からの全ての遷移は等確率となる。ガウシアン出力分布を初期化するために、スピーカーに対するトレーニングデータから全体平均および対角共分散マトリックスが算出される。全ての状態に

対するガウシアン出力分布についての共分散マトリックスが全体的共分散マトリックスに設定される。全体平均に小さな定数を加えることによって平均が設定され、その場合該定数は異なる各状態に対するランダム要素に対して加えられる。Baum-Welch繰り返し処理がスピーカーのトレーニングデータを用いて次に実行される。

【0025】認識されるスピーカーが事前にわかっている場合、Baum-Welchアルゴリズムに対するトレーニングデータは、30秒から1分の各スピーカーに対する音声データを使用して得られる。音声はスピーカーの通常の発声スタイルを表さなければならないが、この場合使用される実際の単語は重要でない。

【0026】スピーカーおよびサイレンスサブネットワークに加えて、ガーベッジ(garbage)サブネットワークが頻繁に使用され、スピーカーモデルまたは存在可能な非音声の内の1つによって特定されない任意のスピーカーがモデル化される。ガーベッジネットワークの形態は、図6に示されるスピーカーネットワークのそれと同じである。しかし、アプリケーションに依存してガーベッジネットワークは異なるデータを使用してトレーニングされる。例えば、ガーベッジサブネットワークが使用されて非音声音がモデル化される場合、それはスピーカーモデルとしてトレーニングされなければならないが、この場合非音声データが使用される。システムに対して未知のスピーカーをモデル化する場合、トレーニングデータを得る1つの方法は、既知の各スピーカーからの音声の部分を使用することである。

【0027】ガーベッジモデルをトレーニングする際に全てのスピーカーからの全てのデータが必ずしも使用されないことは重要である。全ての有効なデータを使用することは、各スピーカーモデルに対してよりもガーベッジモデルに対してより多くのトレーニングデータを与え、全てのスピーカーに対してより確実なスピーカーモデルを作成する効果を有する。従って、結果として得られるHMMネットワークはほとんどの音声をガーベッジとして分類する。

【0028】1実施例において、入力オーディオトレーニングデータは8KHzでサンプルされ、10msごとに特徴ベクトルが算出される。例えば、各フレームに対する特徴ベクトルは、25msウィンドウ下のサンプルに関する20次の線型予測符号化(LPC)を行うことによって算出されることが可能であり、従ってLPCスペクトルから20個のケプストラム定数が算出されることが可能である。

【0029】いくつかの場合においては、認識されるスピーカーは事前にわかっていない。しかし、スピーカーモデルに対する初期推定を得ることがそのような場合にも必要である。この初期推定は、階層的な集塊性のクラスタリングを使用して行われ、異なるスピーカーとして認識されるデータのラフな区分が作成される。

【0030】スピーカに從うデータの区分を与えることによってスピーカサブネットワークの初期推定を得るために、階層的な集塊性のクラスタリングが使用されることが可能である。このデータは次にスピーカHMMのBaum-Welchトレーニングに対するトレーニングデータとして使用されることが可能である。

【0031】セグメンテーションされていないデータは、最初に等しい長さのセグメントに分割され、各セグメントは数秒の音声から成る。これらのセグメントは階層的クラスタリングに対する初期クラスタ集合として使用される。該アルゴリズムは、最初に全てのクラスタペアについてのクラスタ間距離を算出し、次に最も近い2つのクラスタを併合することによって進行する。このプロセスは所望のスピーカクラスタ数が得られるまで繰り返される。このプロセスが図8に概略的に示されている。スピーカ数が未知の場合、このアルゴリズムが使用されてスピーカ数が推定されることが可能である。その場合、最近接クラスタの併合は、最近接クラスタ間距離が定められたスレシヨルドを越えるまで継続する。スレシヨルドを越えるとクラスタリングは中止され、その時のクラスタ数がスピーカ数の推定値として使用される。

【0032】図8は、スピーカでラベル付けされているインターバル集合上の階層的クラスタリング100を概略的に示す。オリジナルインターバル102は、C、L、およびTで3つのスピーカに対してラベル付けされたツリーのリーフによって示される。そのような全てのインターバルについてのインターバル間距離が算出され、104に示されるように最も近接する2つのインターバルが併合される。

【0033】この最近接クラスタ併合プロセスは、所望のクラスタ数が形成されるまで繰り返される。3つのクラスタに対し、それらクラスタに対応する3つの分岐が示されている。第1のクラスタ106はほとんどスピーカCからのインターバルを含み、第2のクラスタ108はほとんどスピーカLからのインターバルを含み、第3のクラスタ110はほとんどスピーカTからのインターバルを含む。

【0034】スピーカ数が未知の場合、距離に対するスレシヨルドが設定され、スレシヨルドが越えられた場合にクラスタの併合が中止される。このことは線112により概略的に示されており、該線は4つのクラスタを生成する。(クラスタ1は2つに分割されている。) クラスタXが単一セグメント $X = x$ かまたはセグメント集合 $X = x_1, x_2, \dots$ から成ると仮定する。クラスタXおよびY間の距離は $d(X, Y)$ により表される。前述のシステムにおいて、セグメント間距離はガウシアン分布の仮定に基づき尤度比によって導出された。 $x = s_1, \dots, s_r$ はある1つのセグメント内のデータを表し、 $y = s_{r+1}, \dots, s_n$ はその他のセグメント内のデー

タを表し、 $z = s_1, \dots, s_n$ は合成セグメント内のデータを表すものとする。 $L(x, \theta_x)$ は x シーケンスの尤度とし、ここで θ_x はガウシアン分布のパラメータに対する推定値である。同様に $L(y, \theta_y)$ は y シーケンスの尤度とし、 $L(z, \theta_z)$ は合成シーケンス z の尤度とする。 λ は尤度比を表すとすると、次式のように表される。

【0035】

【数1】

$$\lambda = \frac{L(z, \theta_z)}{L(x, \theta_x) L(y, \theta_y)}$$

【0036】クラスタリングの際に使用される距離計量は $-\log(\lambda)$ である。音声データは単一のガウシアン分布では十分にモデル化されないため、尤度比はガウシアン分布の混成結合に拡張される。セグメンテーションされていないデータが最初に使用され、M個のガウシアン分布の混成に対する平均および共分散マトリックスが推定される。次にこれらは残りの解析により確定される。 $N_i(s) = N(s; \mu_i, \sigma_i)$ は i 番目の混成要素に関連するガウシアン分布とし、 $g_i(x)$ はデータシーケンス x を使用して推定された i 番目の混成要素に対する重みとする。 $g_i(x)$ は $N_i(s)$ が最大となる x 内のサンプルの割合である。従って x シーケンスの尤度は次式のように表される。

【0037】

【数2】

$$L(x, \theta_x) = \prod_{j=1}^r \sum_{i=1}^M g_i(x) N_i(s_j)$$

【0038】ここで $\theta_x = (g_1(x), \dots, g_M(x))$ である。尤度 $L(y, \theta_y)$ も同様に算出される。合成シーケンスに対する尤度 $L(z, \theta_z)$ の算出において、混成要素に対する重み $g_i(z)$ として次式を得る。

【0039】

【数3】

$$g_i(z) = (r/n) g_i(x) + ((n-r)/n) g_i(y)$$

【0040】クラスタリングに対する距離計量、 $d_i = -\log(\lambda_i)$ は従って式(1)を使用して算出されることが可能である。

【0041】本発明のクラスタリングプロシジャは、クラスタを含むインターバルにおけるインターバル間距離の最大、最小、もしくは平均を使用するよりもむしろ式(1)を使用して集塊性のクラスタ間距離を再計算する点において、通常の階層的クラスタリングと異なっている。従って式(2)および(3)により与えられる尤度の計算効率が重要となる。これはクラスタリングレベルの各々において距離が再計算されるためである。

11

【0042】加えて、スピーカーチェンジの事前確率はM個のスピーカーを伴うマルコフデュレーションモデルを使用して算出されることが可能である。 S_i はセグメントiの期間中のスピーカーを表し、Mはスピーカー数を表すとする。 S_i は、各スピーカーaに対して $P_r [S_{i+1} = a | S_i = a] = p$ 、および各スピーカーaおよびb (aに等しくない) に対して $P_r [S_{i+1} = b | S_i = a] = (1-p) / (M-1)$ を伴うマルコフ連鎖であると仮定する。セグメントiに対するスピーカーがセグメントi+nに対しても発声する確率 $P_r [S_{i+n} = S_i]$ は、2状態マルコフ連鎖を使用して算出されることが可能であり、その場合連鎖の状態1は時刻iにおけるスピーカーを表し、状態2は他の全てのスピーカーを表す。この連鎖に対する遷移確率マトリックスPは次式のように表される。

【0043】

【数4】

$$P = \begin{pmatrix} p & 1-p \\ \frac{(1-p)}{M-1} & 1 - \frac{(1-p)}{M-1} \end{pmatrix}$$

【0044】このマトリックスに関し、 $P_r [S_{i+n} = S_i] = (P^n)_{11}$ である。Pを対角化することにより、 $P_r [S_{i+n} = S_i]$ は次式のようによりクローズした形態で表されることが可能である。

【0045】

【数5】

$$f(n) = P_r [S_{i+n} = S_i] = \frac{1 + (M-1) \left(\frac{Mp-1}{M-1} \right)^n}{M}$$

【0046】この式を使用して、2つの与えられたクラスが同一のスピーカーまたは2つの異なるスピーカーによって生成される事前確率を算出することが可能である。Cをスピーカーチェンジが発生するインターバル数とし、 n_i をi番目のインターバル長とすると、デュレーションバイアスは次式のように定義される。

【0047】

【数6】

$$\lambda_D = \frac{\prod_i^C f(n_i)}{(M-1) \prod_i^C (1 - f(n_i)) / (M-1)}$$

【0048】デュレーションバイアスされた距離は $d_{||}(X, Y) = -\log(\lambda_{||}) - \log(\lambda_{||})$ として定義される。

【0049】図9に示されるスピーカーセグメンテーションネットワーク120は、各スピーカーに対するサブネットワーク60と、サイレンスおよびガーベッジに対するオプションなサブネットワーク64および122とから成る。ガーベッジは、オーディオ中の未知のスピー

12

カーまたは非音声音のような、スピーカーまたはサイレンスモデルによってモデル化されない音声または音として定義される。スピーカー、ガーベッジ、およびサイレンスサブネットワークは以下に述べるように得られる。ネットワークモデルは、2またはそれ以上のスピーカーによるバックグラウンドノイズを伴う会話をモデル化する。

【0050】ネットワーク60のような個々のスピーカーサブネットワークは互いに並列に結合され、各サブネットワークから外部への遷移確率は小さいペナルティ定数 ϵ に固定されて、孤立サンプルに基づくスピーカーチェンジが抑制される。各スピーカーサブネットワーク60はL個の状態を伴うHMMから成り、それらHMMは並列に接続される。各状態は、ガウシアン出力分布、自己遷移、および他状態への遷移を有する。

【0051】初期ヌル状態からスピーカー、ガーベッジ、およびサイレンスサブネットワークへの遷移確率は画一的である。スピーカー、ガーベッジ、およびサイレンスモデルから外部への遷移確率ペナルティは定数 ϵ に設定される。原理的に、これら遷移確率はスピーカーに依存し、トレーニング期間中に学習される。しかし、簡素化を目的として、スピーカーの事前確率は画一的に仮定され、スピーカーを離れる確率 ϵ は経験的に選択されて孤立サンプルに基づくスピーカーチェンジが抑制される。

【0052】実際には、この遷移確率は著しく小さい。

(10^{-20} のオーダーである。) 従って各スピーカーモデルから外部への遷移は、スピーカーからスピーカーへの切替にペナルティを与えるよう作用する。

【0053】スピーカー間の会話をインデックス化することは単に、観測された特徴ベクトルに関する与えられたシーケンスであるネットワークモデルを介する最も確からしい状態シーケンスを見出すことである。スピーカーサブネットワークが初期化された後、スピーカーセグメンテーションネットワークを介する最も確からしい状態シーケンスを見出すことによりスピーカーセグメンテーションが実行され、状態パスがスピーカーを変更する時点でマーキングが施される。最適な状態が1つのスピーカーモデルから他のスピーカーモデルへ切り替わる場合にスピーカーチェンジが発生する。最適な状態シーケンスを見出すことはビタビアルゴリズムを使用して達成される。セグメンテーションの確度は、セグメンテーションされたデータを使用してスピーカーサブネットワークを再トレーニングすることによって改善されることが可能である。このセグメンテーションおよび再トレーニングのプロセスは、セグメンテーションにおいて変化が生じなくなるまで繰り返される。

【0054】部分的トレースバックの方法または連続的デコーディングがビタビ探索で使用される。部分的トレースバックは、Brown らによる「Partial Traceback an

50

d Dynamic Programming」(Proc. Int. Conf. Acoustics, Speech and Signal Processing, May 1992, pp. 1629-1632)に記載されている。このアルゴリズムにおいて、全ての状態からの各タイムステップにおいてビタビトレースバックが実行され、全てのパスの初期部分が整合する場合にはデコーディングが可能となる。実際に被る遅延は1秒未満である。

【0055】後処理のアプリケーションにおいて、音声のセグメンテーションが繰り返し実行されることが可能であり、その場合各セグメンテーションの後にスピーカ
10 モデルが再トレーニングされる。このことはセグメンテーションの精度を向上させ、特にスピーカトレーニングデータが使用不可能な場合に有効である。

【0056】繰り返し再セグメンテーションアルゴリズムが図9に示される。前述のように、最初にトレーニングデータ集合がボックス130のステップで与えられ、ボックス132のステップでスピーカモデルがトレーニングされる。次にボックス134のステップでこれらスピーカモデルに基づきセグメンテーションが実行される。ボックス134のステップでのセグメンテーションが大きく変化する場合、この改善されたセグメンテーションはスピーカに対する新たなトレーニングデータとして使用され、ボックス132のステップでスピーカ
20 モデルが再トレーニングされる。このプロセスはボックス136のステップでセグメンテーションが変化しなくなるまで続けられる。

【0057】図10は、オーディオデータが記録デバイスによって記憶媒体上に記憶される場合にリアルタイムでオーディオデータのインデックスを作成する、本発明のシステム140を示す。

【0058】メモリ148から命令を得るシステムプロセッサ146はトレーニングデータ147を受信してスピーカモデルを決定する。スピーカモデルは結合され、後のオーディオストリーム処理のためのスピーカネットワークが形成される。トレーニングデータ147は、識別される各スピーカに対するトレーニングデータを有していなければならない。図10に示されるように、トレーニングデータ147はそのオリジナルなオーディオ波形から既に処理されており、スペクトル特徴データの形態でシステムプロセッサ146に保存されている。
40

【0059】オーディオ入力141はオーディオプロセッサ142によってスペクトル特徴データへ処理され、システムプロセッサ146に与えられる。これと同時に、オーディオ入力はオーディオ記録デバイス143によって記憶媒体144上に記録される。記録デバイス143は、オーディオストリーム情報をアナログまたはデジタル形態で記憶することが可能であり、純粋なオーディオ記録、もしくはオーディオ/ビデオ記録の部分であることが可能である。
50

【0060】スペクトルデータは、システムプロセッサ146によってトレーニングデータ147から作成されたスピーカネットワークを使用することによってシステムプロセッサ146によって処理される。オーディオストリームにおいて新たなセグメントの各々が検出されると、システムプロセッサ146はタイムソース145からタイムスタンプを得る。タイムスタンプは、オーディオデータの記憶媒体144上への記憶時間を示す。タイムソース145は、例えば、記録が開始される時に始動する時計であることが可能であり、もしくは、記憶媒体に接続された記録デバイスから時間を記録するデバイスであることが可能である。このタイムスタンプは、セグメントの作成者の識別子と共にメモリ148に記憶され、後にスピーカに従うインデックスへ収集される。

【0061】図11は、オーディオ記録データのスピーカに従うインデックスを作成および記憶する、システム190における本発明のその他の実施例を示す。

【0062】トレーニングデータ196はシステムプロセッサ194へ与えられ、スピーカモデルおよびスピーカネットワークが生成される。トレーニングデータ196は識別される各スピーカに対するトレーニングデータを有していなければならない。図11に示されるように、トレーニングデータは、既にそのオリジナルなオーディオ波形から処理されており、スペクトル特徴データとしてシステムプロセッサ194に保存されている。識別された各スピーカに対して記録の部分が孤立され得る場合、トレーニングデータはまたオーディオ記録入力191の部分であることが可能である。

【0063】オーディオ記録入力191はオーディオプロセッサ192によってスペクトル特徴データへ処理され、システムプロセッサ194へ与えられる。スペクトル特徴データは、システムプロセッサ194による後の繰り返し処理のためにメモリ197に記憶されることが可能である。スペクトルデータは、システムプロセッサ194によってトレーニングデータ196から作成されたスピーカネットワークを使用することによってシステムプロセッサ194によって処理される。オーディオストリームにおいて新たなセグメントの各々が検出されると、システムプロセッサ194はタイムソース193からタイムスタンプを得る。タイムスタンプは、オーディオ入力191の記録からのオーディオデータの記録アドレスまたは記憶時間を示す。タイムソース193は、例えば、記録が開始される時に始動する時計であることが可能であり、もしくは、記憶媒体に接続された記録デバイスから時間を記録するデバイスであることが可能である。このタイムスタンプは、セグメントの作成者の識別子と共にメモリ195に記憶され、後にスピーカに従うインデックスへ収集される。

【0064】図10のシステム140によって記録されるオーディオデータは、図12のシステム190におい

て記録データ 191 として使用されることが可能である。そのような場合、システム 140 によって作成されるインデックスが使用されてトレーニングデータ 196 が与えられることが可能であり、特定のスピーカに属するセグメントの各集合は、新たなスピーカモデルをトレーニングするトレーニングデータとして使用される。システムプロセッサ 194 は新たなスピーカモデルを使用し、それらを結合してネットワークとし、オーディオストリームの再セグメンテーションを行う。

【0065】そのような繰り返し処理は、システム 140 からシステム 190 へのものであれ、もしくはシステム 190 を繰り返し介するものであれ、セグメンテーションの精度をさらに向上させる。

【0066】図 12 は、オーディオストリームのインデックスを決定する前述の方法のアプリケーションを示す。図 12 におけるステップはリアルタイムもしくは後処理モードで実行されることが可能である。オーディオストリームは通常、オーディオタイミングに相関付けられたセグメント情報と共に記録される。ボックス 150 のステップは既知のスピーカからトレーニングデータを選択する。前述のように、そのようなトレーニングデータは個々のスピーカによる 30 秒から 1 分の音声から成ることが可能である。このトレーニングデータがボックス 152 のステップで使用されて個々のスピーカ各々に対する HMM スピーカモデルがトレーニングされる。

【0067】ボックス 154 のステップにおいて、図 8 に関連して述べられたように、個々のモデルが並列に接続されてスピーカセグメンテーションネットワークが形成される。この時点で、個々のスピーカモデルから離れることに対するペナルティが挿入される。ボックス 156 のステップはガーベッジ、即ち未知のスピーカおよび／または非音声音、およびサイレンスインターバルに対するモデルを作成および付加する。サイレンスおよびガーベッジモデルはボックス 152 のステップで既に作成されていることも可能である。

【0068】ボックス 158 のステップにおいて、オーディオストリームはスピーカセグメンテーションネットワークを使用してセグメントに分解される。セグメントは、ボックス 160 のステップにおいて各セグメントに対するスピーカの識別子を用いてマーキングされる。ボックス 162 のステップは同様のマーキングが施されたセグメントを収集してオーディオ記録のスピーカインデックスを作成する。

【0069】リアルタイム動作が必要ない場合、図 13 に関して述べられるように、より詳細な処理が実行されることが可能である。ボックス 170 から 180 に示されるステップは、図 12 のボックス 150 から 160 に示されるステップに関して述べられた方法と同様に実行される。

【0070】ボックス 182 に示されるステップにおいて、テストが実行され、ボックス 178 のステップで決定されたセグメンテーションが前のセグメンテーションから変化したかどうか決定される。セグメンテーションに大きな変化があった場合、システムはボックス 172 のステップに戻り、スピーカモデルの再トレーニングおよびオーディオストリームの再セグメンテーションが行われる。シーケンスの初期には前のセグメンテーションが存在しないためシステムは前述のように繰り返しを行う。オーディオストリームのセグメンテーションにおいて繰り返し処理により大きな変化が生じなくなった場合、同様のマーキングが施されたセグメントが収集されてボックス 184 のステップでインデックスが作成されることが可能である。

【0071】情報のインデックス化は、スピーカ ID に基づきオーディオストリームの再生を検索および制御する能力をユーザーに与えることが可能である。例えば、ユーザーは特定のスピーカによる発言のみを検索したいかもしれない。ユーザーはさらに、オーディオインデックス情報を使用してオーディオ記録を検索するかもしれない。ユーザーは、いくつかのスピーカセグメントをスキップし、次のスピーカへ効果的に早送りを行うかもしくは特定のスピーカセグメントの始まりへ巻戻しを行いたいかもしれない。

【0072】

【発明の効果】以上説明したように、本発明の方法によれば、オーディオストリーム内にインデックスを作成することが可能となり、リアルタイムであっても処理後であっても、ユーザーが特定のスピーカに関連するオーディオデータセグメントを認識することが可能となる。

【図面の簡単な説明】

【図 1】本発明が実施されることが可能である一般化されたオーディオ処理システムのブロック図である。

【図 2】オーディオインデックスシステムの一般化されたフロー図である。

【図 3】5 状態隠れマルコフモデル (HMM) を示す図である。

【図 4】HMM によってモデル化される 3 つの対象物の HMM ネットワークを示す図である。

【図 5】ビタビアルゴリズムの結果を概略的に示す図である。

【図 6】個々のスピーカの発声スタイルをモデル化する 3 状態 HMM を示す図である。

【図 7】サイレンスサブネットワークを示す図である。

【図 8】各スピーカに対するサブネットワークと、サイレンスおよびガーベッジに対するオプションなサブネットワークとから成るスピーカセグメンテーションネットワークを示す図である。

【図 9】繰り返し再セグメンテーションアルゴリズムを概略的に示す図である。

17

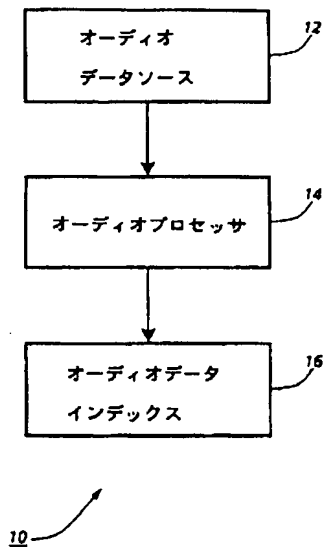
【図10】オーディオデータが記録デバイスによって記憶媒体上へ記憶される場合にリアルタイムにオーディオデータのインデックスを作成する本発明に従うシステムを示す図である。

【図11】事前に記録されたオーディオデータのスピーカーに従うインデックスを作成および記憶するシステムにおける本発明のその他の実施例を示す図である。

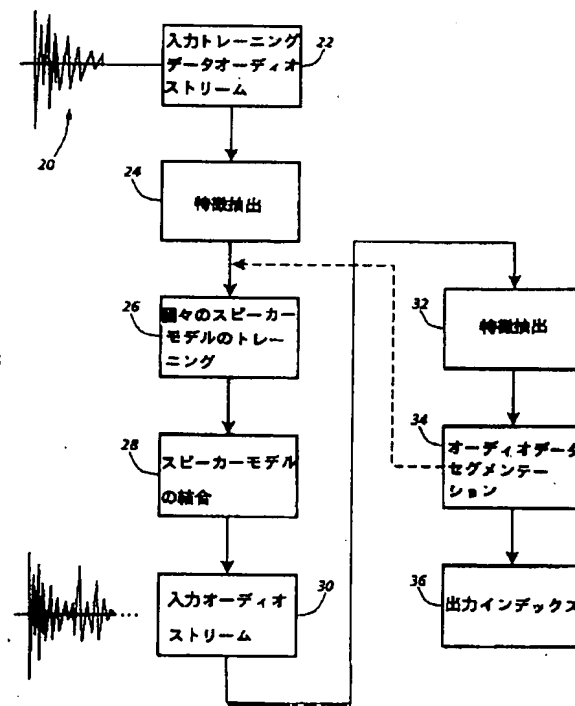
【図12】スピーカーが未知の場合にオーディオストリームのインデックスを決定する本発明に従う方法を示す図である。

10

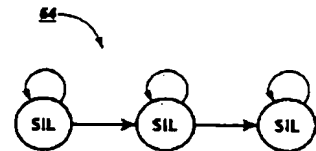
【図1】



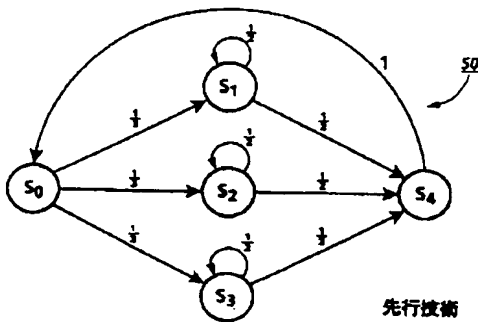
【図2】



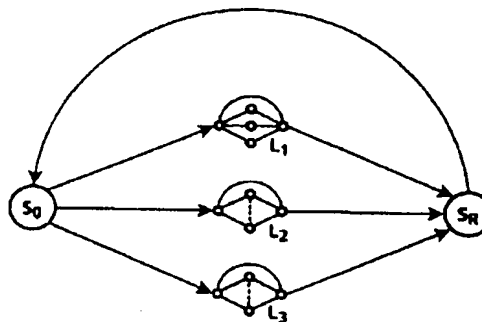
【図7】



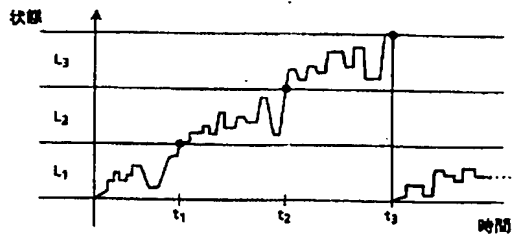
【図3】



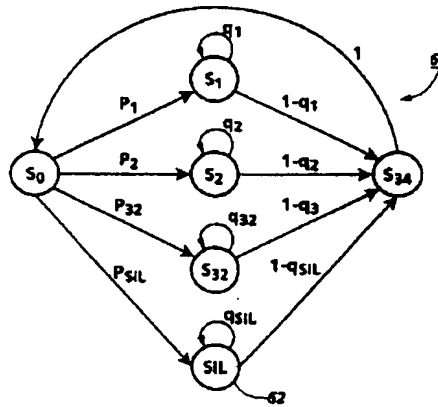
【図4】



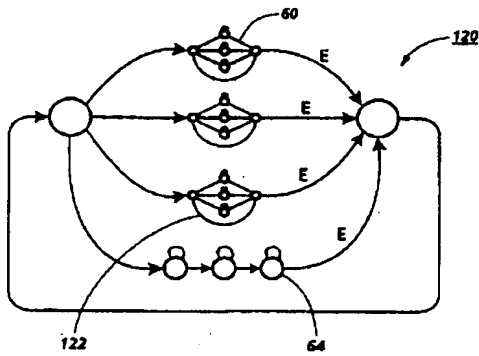
【図5】



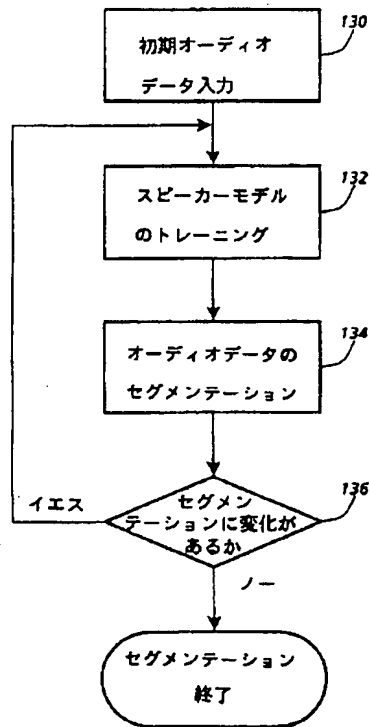
【図6】



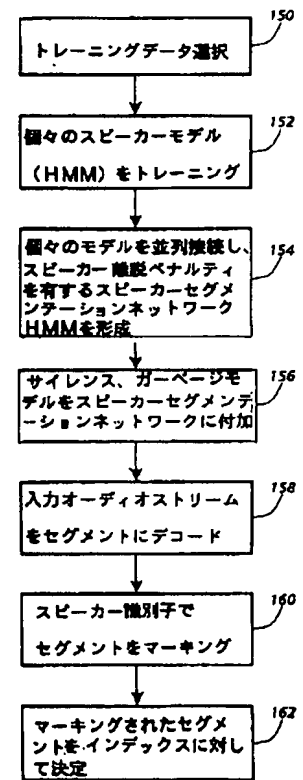
【図8】



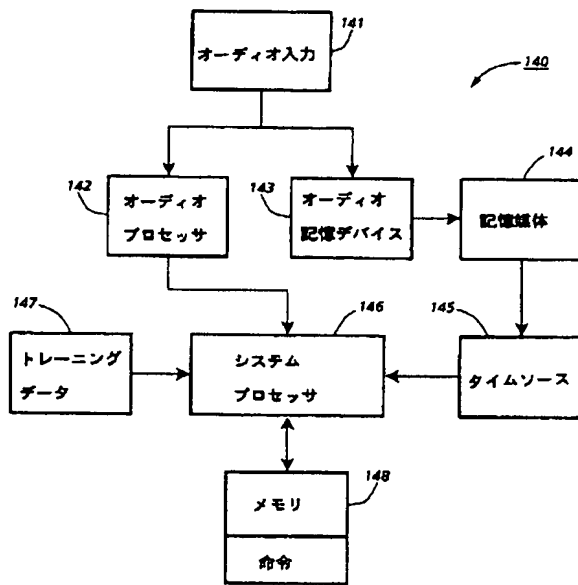
【図9】



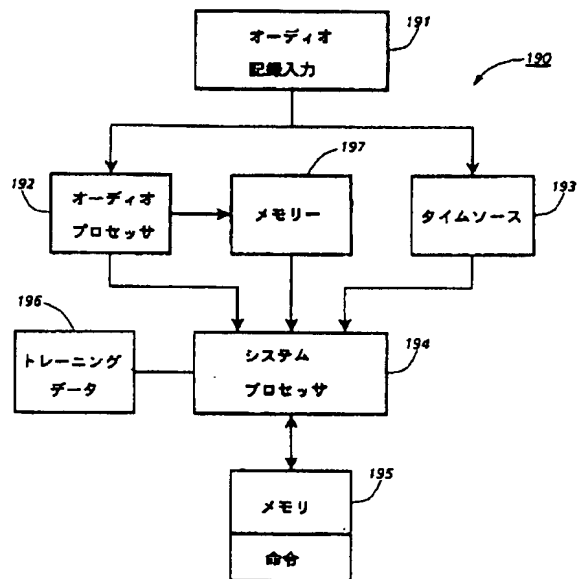
【図12】



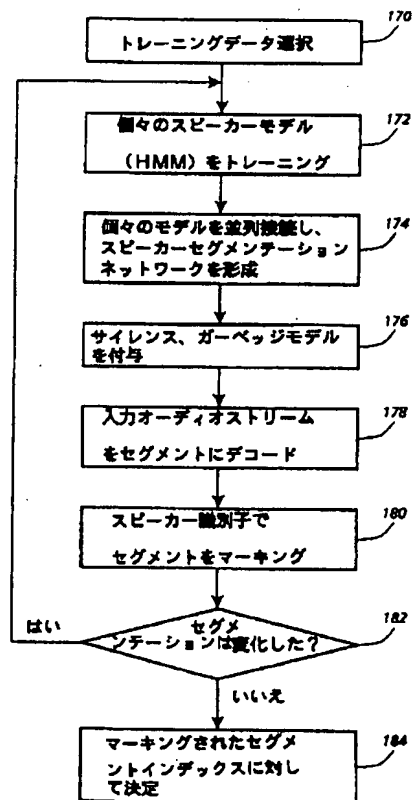
【図10】



【図11】



【図13】



フロントページの続き

- (72)発明者 フランシン アール. チェン
アメリカ合衆国 カリフォルニア州
94025 メンロ パーク シャーマン ア
ヴェニュー 975
- (72)発明者 フィリップ エイ. チョウ
アメリカ合衆国 カリフォルニア州
94025 メンロ パーク ブラックバーン
アヴェニュー 116
- (72)発明者 ドナルド ジー. キンバー
アメリカ合衆国 カリフォルニア州
94040 マウント ビュー ヴィクター
ストリート 678 ナンバー 3

- (72)発明者 アレックス ディー. プーン
アメリカ合衆国 カリフォルニア州
94040 マウンテン ビュー サウス レ
ングストーフ アヴェニュー 575 アパ
ートメント ナンバー 21
- (72)発明者 カロン エイ. ウェバー
アメリカ合衆国 カリフォルニア州
94109 サンフランシスコ ユニオン ス
トリート 1330 ナンバー 22
- (72)発明者 リン ディー. ウィルコックス
アメリカ合衆国 カリフォルニア州
94028 ポートラ ヴァレー ジョアクイ
ン ロード 45